DOCUMENT RESUME

ED 088 929 TM 003 489

AUTHOR Fennessey, James
TITLE Understanding "Fan-Spread" in Achievement Measures.
INSTITUTION Johns Hopkins Univ., Baltimore, Md. Center for the Study of Social Organization of Schools.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO R-168
PUB DATE Jan 74
CONTRACT NE-C-00-3-0114
NOTE 29p.

EDRS PRICE MF-$0.75 HC-$1.85
DESCRIPTORS *Achievement Tests; *Evaluation; *High Achievers; Longitudinal Studies; *Measurement Techniques; Models; *Scores; Test Interpretation

ABSTRACT

If educators hope to use educational achievement test scores to assess the impact of instructional programs, they must deal with the phenomenon of "fan-spread"--the tendency for groups with higher average scores to have a greater amount of within-group dispersion of scores. This paper discusses the nature of fan-spread, how it appears in measurement situations, and its complexity in real data. The final section generalizes to the question of choosing a score format for analytical use. (Author)

Center for Social Organization of Schools

REPORT NO. 168
January, 1974

UNDERSTANDING "FAN-SPREAD" IN ACHIEVEMENT MEASURES

James Fennessey

STAFF

John L. Holland, Director

James M. McPartland, Assistant Director

Joan E. Brown

Judith P. Clark

David L. DeVries

Joyce L. Epstein

James J. Fennessey

Ann Forthuber

Stephanie G. Freeman

Ellen Greenberger

Edward J. Harsch

John H. Hollifield

Ruthellen Josselson

Nancy L. Karweit

Daniel D. McConochie

Edward McDill

James W. Michaels

Dean H. Nafziger

James M. Richards

John P. Snyder

Julian C. Stanley

UNDERSTANDING "FAN-SPREAD" IN ACHIEVEMENT MEASURES

CONTRACT NO. NE-C-00-3-0114

WORK UNIT NO. 2A

JAMES FENNESSEY

REPORT NO. 168
JANUARY, 1974

Published by the Center for Social Organization of Schools, supported in part as a research and development center by funds from the United States National Institute of Education, Department of Health, Education and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the Institute should be referred.

The Johns Hopkins University

Baltimore, Maryland

Introductory Statement

The Center for Social Organization of Schools has two primary objectives: to develop a scientific knowledge of how schools affect their students, and to use this knowledge to develop better school practices and organization.

The Center works through three programs to achieve its objectives. The Schools and Maturity program is studying the effects of school, family, and peer group experiences on the development of attitudes consistent with psychosocial maturity. The objectives are to formulate, assess, and research important educational goals other than traditional academic achievement. The School Organization program is currently concerned with authority-control structures, task structures, reward systems, and peer group processes in schools. The Careers and Curricula program bases its work upon a theory of career development. It has developed a self-administered vocational guidance device and a self-directed career program to promote vocational development and to foster satisfying curricular decisions for high school, college, and adult populations.

This report, prepared by the School Organization program, examines the phenomenon of "fan-spread" that appears in educational achievement measurement.

Abstract

If educators hope to use educational achievement test scores to assess the impact of instructional programs, they must deal with the phenomenon of "fan-spread" - the tendency for groups with higher average scores to have a greater amount of within-group dispersion of scores. This paper discusses the nature of fan-spread, how it appears in measurement situations, and its complexity in real data. The final section generalizes to the question of choosing a score format for analytical use.

Understanding "Fan-Spread" in Achievement Measures

This paper examines "fan-spread," a feature of educational achievement tests that often proves troublesome to those who try to use such scores to assess the impact of instructional programs. Fan-spread refers to the fact that if the achievement score distributions of several groups are examined, those groups whose average score is higher also tend to have a greater amount of within-group dispersion of scores. The phenomenon is most widely recognized in standardized achievement test data, of the sort now collected by a large number of school districts. It can and does appear, however, in other kinds of achievement measures.

In any real situation, the observed test scores are the result of a complex combination of processes, but this paper will begin with a simplified example. The first portion of the paper will describe a measurement situation which appears quite different from that of achievement testing, but which is formally quite similar. In both of these measurement situations, fan-spread is a pattern which may or may not occur. The next section of the paper will present in fairly simple algebra, and with an artificial numerical example, a measurement situation with fan-spread.

The next two sections of the paper will re-introduce some of the practical complexity that had been temporarily neglected. In real data, fan-spread is neither simple nor constant. Moreover, it occurs in combination with other artifactual influences upon the test score-- such as regression effects-- and thus makes interpretation of scores very problematic. Following that, another section will describe an adjustment approach of use in situations where the fan-spread dynamics are pure.

The last section of the paper will generalize from the preceding ones to the question of choosing a score format for analytical use. This will be discussed in terms of the general logic of model-building and utilization; some general recommendations will be made.

The Analogy of the Military Obstacle Course

Fan-spread may well occur not only in achievement tests or other kinds of psychometric measures which are scaled against inferred continuums, but also in a parallel situation where the scale of measurement is that of ordinary physical distance. There is essentially no doubt about the general utility of the ordinary scales for distance, so the fan-spread phenomenon may be more objectively considered in that context.

Suppose, for example, one was interested in describing the performance of several groups of soldiers on a military obstacle course. To describe the level of performance of an individual soldier, it would be reasonable to establish a benchmark time interval and then record the distance on the course covered by the soldier in that time interval. Let us consider the parallelism between the obstacle course and the school achievement situation. The parallel between the distance covered on the obstacle course and the amount of material learned by a student is quite evident. The problems in recognizing the similarity arise because the specific scale used for distance is independently and firmly established by direct validation. This is not possible with scales for academic achievement, so the exact nature of the achievement scale is somewhat debatable. However, it is possible to suppose that a particular scale of achievement is a good approximation of a linear scale of distance, and then to consider whether any of the patterns in the data force us to reject that supposition.

To explore the parallel between the obstacle course and the achievement test, consider some of the factors which might be expected to influence the distance covered by an individual soldier on the course in a fixed time. This distance would depend upon (1) the length of the time interval; (2) the average level of difficulty of the course; (3) the distribution of particularly difficult parts over the entire length; (4) the general physical condition and strength of the individual soldier; and (5) the fact that a given soldier might be completely stymied by one particular obstacle, even though he could easily master the others. Other factors could be suggested, but the above would undeniably be important.

More concretely, one might conceptualize the average difficulty as simply the fact that the course is constructed on the side of a hill, so that the soldiers always are progressing uphill. Then, if there are several distinct courses, each characterized by a distinct constant slope, we could use the steepness of slope as a measure of the course's difficulty. This measure would be logically independent of the performance of any soldier upon the course.

The analogy between the obstacle course and the achievement test is close, but has some limitations. The most important one is that the performance on the obstacle course is what the soldier is trained in and it is what he will be expected to perform when in combat. For the educational achievement test (at least for the usual commercially-published, nationally normed test) the relationship between the test and the training activities, and between the test and subsequent performance dimensions, is far less close. This less close linkage between test activities and either training or performance further complicates the decision as to which particular summary measure is most appropriate.

The military obstacle course is much more similar to the currently fashionable notion of the "criterion-referenced" test, and "behavioral objectives".

After agreeing that there is a close analogy, in measurement terms, between the military obstacle course and the educational achievement test, the next question is whether this analogy can lead to insight about the patterns of scores on educational tests, and particularly about the fan-spread pattern. For example, the question can be asked, "would one expect to find the fan-spread pattern in the scores of several groups of soldiers on an obstacle course?" Consider what that question might mean. There are several groups of soldiers, and all of the soldiers in each group have gone through the obstacle course. The distance covered by each soldier in the established time interval has been recorded as that soldier's score. We know that the average score of soldiers in some of these groups is considerably higher than the average score of soldiers in other groups. The question of fan-spread might be phrased as follows: is there a positive correlation between the mean score of a group and the standard deviation of scores in that group?

It is imaginable that such a correlation might be found. One can think of at least three ways in which such a pattern might have arisen. First, the groups might have been allowed differing lengths of time to work on the course. The possibility is excluded in our example, but would have to be considered in a more general case. Second, the fan-spread phenomeonon might be expected if the average difficulty of the obstacle course for some of the groups was different from that for other groups. This too is ruled out in our example, but would be possible in general. The third possibility would be that the general ability of the soldiers in some of the groups was higher than the general ability of the soldiers in other groups. Under those conditions, which <u>are</u> possible in our example, the average distances

covered would diverge. Suppose now that in fact the soldiers were placed into groups strictly according to a rank order of their initial ability on the obstacle course. It is quite possible, though not necessary, that the dispersion of initial ability levels in the top group would be greater in magnitude than the dispersion of initial abilities in the lower groups. This pattern of ability distributions could produce fan-spread within the groups in the same way that the dispersion of abilities between groups can produce fan-spread between groups.

This line of reasoning, including the two possible explanations of fan-spread which are not allowed in the example, is based on an abstract model in which the amount of distance covered per unit of time is proportional to an individual's ability and the difficulty of the material. This model is reasonable, both for the obstacle course and for academic learning. The important complication brought out in the above discussion is the fact that the group average performance and the performance of individuals are related only in a complex way. To anticipate slightly the argument, this relationship is based on that for decomposition of an observed average dependent variable into between-group and within-group components, as is done in analysis of variance, or more generally in analysis of covariance. The next section of the paper seeks to present these ideas more concretely and precisely through use of an artificial example.

Before turning to a consideration of that example, one other point deserves mention. Although it is quite important, it will not be discussed in detail here, for reasons of length and because of the limited focus of this paper. This point is that in actual situations, the distribution of abilities among group members at a point in time tends to depend to a considerable degree not only upon the previous distribution of these same

abilities, but also upon the policies of the training program for allocating time and training energy. Thus, for example, in some situations, students (or soldiers) who manifest initially poor performance-- which presumably is indicative of poor initial ability-- receive extra attention so that the ability (to the extent that it can be improved) is improved. This kind of a policy will tend to decrease the variability of performance within a group, but perhaps to increase the variability between groups, because of the relative lack of attention to the high initial performers. It seems obvious that different training programs will differ in their policies and actual practices in allocating attention to students/soldiers with various levels of initial performance. Also, a person's ability level can be more or less elastic, depending upon the influences which affect it. If the influences are fairly easily modified, then the performance level is more easily modified. To the extent that the influences are not modifiable for an individual (as would be the case if the influences were genetic), the level of performance will be inelastic.

A Hypothetical Example

Let us consider a situation in which there are two groups of students/soldiers, each containing five students/soldiers. For simplicity, we shall refer to the learners as students. We also assume that the students are placed into groups according to their rank on initial performance level (prior to any training) and that the initial levels are as follows:

High Group: 70,65,60,55,50

Low Group: 50,45,40,35,30

For convenience, we can think of the task as being the obstacle course, but it should be obvious that the analogous application to the educational achievement tests applies at each instance. Assume the the established time

interval is such that in one unit trial, each student will progress exactly as many units of distance as he has units of initial ability. This particular choice of units in no way restricts the generalization.

With these assumptions, we can calculate the average scores and the standard deviations for each of the two groups after one, two, and three time intervals. These results are (using 5 as the divisor for calculation of the variance):

	interval 1		interval 2		interval 3	
	$\bar{X}$	sigma	$\bar{X}$	sigma	$\bar{X}$	sigma
High Group	60	7.071	120	14.142	180	21.213
Low Group	40	7.071	80	14.142	120	21.213

Several points should be noted about these results. First, if one compares the gap between the two groups at the end of the second interval with the same gap at the end of the first interval, the first impression is that the gap has increased. That is, at the end of the second interval the gap is 120 - 80 = 40, but at the end of the first interval the gap is only 60 - 40 = 20. The question here is not the "reality" of this gap in some metaphysical sense, because it clearly does change in size. Instead, the question is how the gap's change should be interpreted. In this example, the change in the size of the gap cannot be attributed to any particular experimental treatment, because no such treatment is involved.

The second point about the pattern of these artificial numbers is that the size of the standard deviation for a given group of students is exactly proportional to the average score for those same students. That is, for the High Group, the standard deviation increases from the end of interval 1 through the end of interval 2 and again through the end of interval 3. For each time point, the standard deviation is 0.118 of the corresponding mean. A similar pattern holds for the Low Group, except that the ratio of the

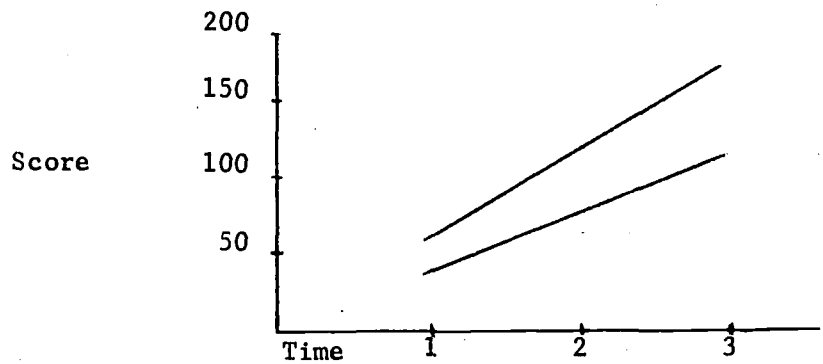
standard deviation to the mean is 0.177.

This example leaves a number of topics neglected. One is that the High and Low groups have exactly the same size standard deviation at each point in time. This is sometimes but not usually found in actual achievement test data. However, the reasons for this difference between the example and the more typical real data lie in some of the special circumstances (such as policy efforts to close the gap at the bottom, and also artifacts of the measuring devices, such as floor effects) which affect the real data. The pattern in this artifical data can be described by saying that there is fan-spread within each group over time (the standard deviations increase over time for each group); there is fan-spread between groups over time (the gap between the means of the two groups increases over time); but there is no fan-spread between groups at a given point in time (the two standard deviations at a point in time are equal). In later sections of the paper, we shall return to this example as a means of considering some of the ways in which actual data are more complicated than the illustration given here.

The example is deliberately simplified and artificial. Its major purpose, however, is to illustrate that even in the situation where the measurement scale is regarded as unchallengable (ordinary distance), and where no treatment whatever exists, a pattern of fan-spread can be expected if the process determining the actual score at a point in time is a cumulative function of exposure time, difficulty of material, and level of individual ability. Because these general properties apply to actual achievement data as well, it seems fair to conclude that fan-spread in actual data also may be (entirely or partly) irrelevant to the interpretation of real differences in the impacts of some program.

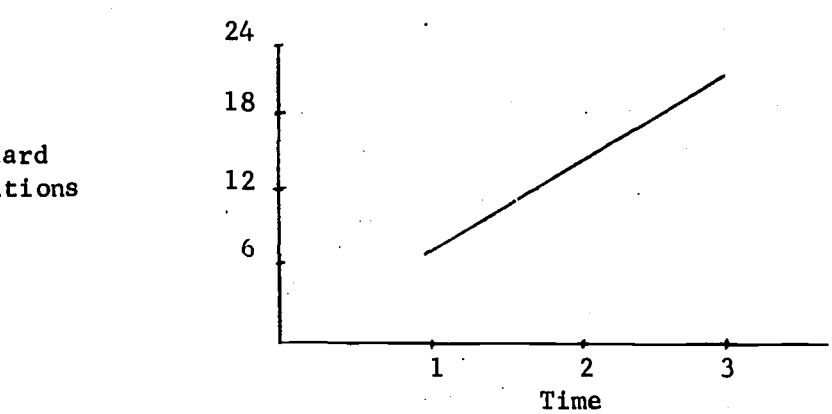
Patterns of Fan-Spread in Actual Achievement Test Data

The fan-spread phenomenon can be represented in a variety of ways. For instance, one can simply present a time plot of the scores of several individuals (or groups) which differ in initial ability. Using the artifical data of the earlier example, the time plot of averages would be:



This plot indicates why Donald Campbell (1971) chose to name this the "fan-spread" phenomenon. This particular plot, or the corresponding tabulation, is an appropriate way to present achievement test data if interest centers on a comparison of the growth rates, and if the interpretation of the fan-spread is not problematic.

A somewhat different approach, also a graphic one, would be to plot the size of the standard deviations as a function of time. This approach, for the artificial data of the earlier example, yields:



This approach is more relevant when interest is on the pattern of fan-spread itself, and one is attempting to examine the dispersion of the standard deviations of various influences. If the concern is to ascertain whether there is a sizeable amount of fan-spread in some batch of achievement test data, then a graph or tabuation of the standard deviations against time seems the most straightforward and direct approach.

Now we can turn our attention to fan-spread possibilities in actual test score data. One set of real data on any particular published test is the set of data on which the test's national norms were based. For most tests, such data actually are obtained by cross-section surveys of grade-cohorts, and not from longitudinal measures of the same students. Also, the data are processed in complex ways to derive the scores used in ordinary reports-- grade equivalents, stanines, percentiles, growth scores, etc. However, for our purposes, the norming study data can be regarded as if they had been obtained from a single series of observations on the same group of students.

Let us consider first some of the normative data for a widely used test battery, the Metropolitan Achievement Tests, 1970 edition. Table 1 presents the standard deviations which were observed in the national norm group for two of the most widely used subscales of this battery, the Total Reading score and the Total Mathematics score. The results in this table are for scores in the "grade-equivalent" format. Other score formats also are fairly often used. In particular, some publishers provide a set of special scores which have been calculated by an elaborate procedure whose aim is to make the scores an approximately equal-interval scale of the underlying trait. For the Metropolitan series, the publisher has provided such scores, and has labeled them "standard" scores. It should be made

Table 1

Standard Deviations for Grade-Equivalent Scores,
Metropolitan Achievement Test, national norm data

Time in Years	Std. Dev. for Total Reading		Std. Dev. for Total Math	
1.7	0.61	0.690	0.71	0.725
2.1	0.77		0.74	
2.7	1.01		0.86	
3.1	1.17		0.92	
3.7	1.38		1.11	
4.1	1.61		1.19	
4.7	1.67		1.35	
5.1	1.83		1.29	
5.7	1.88		1.51	
6.1	2.07	2.020	1.68	1.725
6.7	1.97		1.77	
ratio of smallest pair to largest pair is	2.92		2.38	

clear that in this context, these scores are simply "special"; they are not standard scores in the ordinary sense of the word, but instead are essentially the same as the scores called "growth scores" by S.R.A. (1969). Table 2 presents the standard deviations for the norm group in terms of these publisher's standard scores, for the same two subtests.

Table 1 shows that there is a consistent increase over time in the size of the standard deviation of the grade-equivalent scores for reading and math. For the publisher's standard scores, however, this tendency toward fan-spread is much less evident. For the standard scores in Math (see Table 2) there is no consistent fan-spread operating. This lack of parallellism between the reading and math subtests stands as a puzzle, and one which quickly leads to real-life complexities and ambiguities.

The discrepancy between the two subtests in the amount of fan-spread suggests that there are some general differences between the processes through which reading is taught and learned and those through which mathematics is taught and learned. Measured skills in mathematics seem more easily detected, and shortcomings more evident also -- to the student or to a teacher or other adult helper. Also, because the basic knowledge and skills in math are more algorithmic in structure than the corresponding skills for reading, it is easier for specific difficulties in mathematics to be remedied by a brief lesson or summarized in a brief rule of procedure. This difference in the process of learning may tend to raise the floor of performance on a mathematics test above that of the related tests on reading.

Obviously, this is a conjectural analysis. It suggests, however, some of the benefits and dilemmas that may arise from a careful and detached study of achievement data files.

A related point concerns the possible "treatment" differences in the

Table 2

Standard Deviations for Publisher's "Standard" Scores,
Metropolitan Achievement Test, national norm data

Time in Years	Std. Dev. for Total Reading	Std. Dev. for Total Math
1.7	10.0 } 10.45	12.6 } 12.35
2.1	10.9 }	12.1 }
2.7	10.9	11.1
3.1	11.6	11.4
3.7	13.0	12.0
4.1	14.0	12.0
4.7	14.3	12.1
5.1	13.5	10.4
5.7	13.0	12.2
6.1	14.7 } 14.10	12.1 } 12.40
6.7	13.5 }	12.7 }
ratio of smallest pair to largest pair is	1.35	1.00

national norm data. To the extent that fan-spread is found in the norm data, it may be due only to the process of cumulative-advantages-of-ability which were described in the artificial example. There is, on the other hand, at least one other hypothesis which appears to be worth considering. This hypothesis might be stated as "them that has, gets." In some circumstances, attention given to high-performing students may be greater than the attention given to low-performing students, which would tend to accentuate the pattern of fan-spread. It should be emphasized that no prejudice in favor of high-performing students is being assumed here. Such prejudice is possible, of course, but a number of other mechanisms could and do operate. To some extent, for instance, the community resources available in communities where most of the children are high-performers tend to be larger than in communities where most of the children are low-performers. Another possibility is that low-performing children also have additional needs of a social or emotional nature, so some of the total resources of the school must be devoted to dealing with those needs instead of to direct training on academic skills.

In short, once we move from considering simplified artifical data to real data in real situations, the problems of interpretation become far more complex. Issues of widely different kinds come to be tangled together in the data, and considerable caution (as well as additional data) is needed if practical inferences are to be drawn properly.

To pursue this general point slightly further, and also to illustrate some of the possible uses of the fan-spread aspect of test scores, Table 3 presents some specific real data showing patterns of fan-spread.

The data in Table 3 are taken from a recent field experimental study aimed at assessing the impact of monetary incentives payments as a way to improve academic achievement among low-income students in elementary schools (Planar Corporation, 1972, Appendix B). In this study, standardized

Table 3

Standard Deviations of "Standard" Scores, Metropolitan Achievement Tests
U.S.O.E. Incentives Project (Pretest Only, Data by Grade, School, Site, Subtest)

				1	2	3	4	5	6
1	READ	CIN	EXP	4.53	6.33	10.25	9.25	11.79	12.59
2	READ	CIN	CON	3.57	8.38	7.38	9.04	8.42	12.06
3	READ	JAX	EXP	4.83	6.15	7.77	12.12	11.44	14.48
4	READ	JAX	CON	5.16	6.38	9.62	9.74	10.09	12.08
5	READ	OAK	EXP	4.90	8.62	8.39	12.41	10.74	15.22
6	READ	OAK	CON	4.12	8.43	9.42	9.40	12.17	10.43
7	READ	SAN	EXP	4.11	5.20	10.93	9.57	13.27	12.39
8	READ	SAN	CON	4.69	5.15	9.76	11.39	13.70	9.82
9	MATH	CIN	EXP	6.32	5.39	13.57	12.17	11.35	11.30
10	MATH	CIN	CON	5.22	9.72	10.49	10.30	8.05	10.30
11	MATH	JAX	EXP	5.47	7.82	9.12	12.05	12.46	14.33
12	MATH	JAX	CON	6.37	7.88	9.96	11.43	12.76	12.54
13	MATH	OAK	EXP	4.92	9.17	8.56	14.50	12.13	13.33
14	MATH	OAK	CON	4.67	9.20	11.35	12.31	13.52	12.07
15	MATH	SAN	EXP	6.76	4.84	12.57	11.72	11.87	14.34
16	MATH	SAN	CON	6.16	6.49	11.08	12.32	11.77	9.82
Average of 8-reading				4.4888	6.8300	9.1900	10.3650	11.4525	12.3838
Average of 8-math				5.7363	7.5638	10.8375	12.1000	11.7388	12.2538

achievement tests were administered at the beginning and again at the end of the academic year during which the incentives program was in effect. Because our interest here is only in patterns of fan-spread in ordinary school situations, we have examined only the scores and standard deviations in the pretest data. Thus, the standard deviations presented in Table 3 can be regarded as an ordinary set of test scores on students from grades one through six in eight different schools, two schools per city for each of four cities; all of the schools attended primarily by disadvantaged children.

Table 3 shows the standard deviations by grade and by school, city, and subtest content. The test used was the Metropolitan Achievement Test. Here, as in Table 2, the score format is that for the publisher's "standard" scores -- the specially calculated scores which are presented as being more appropriate for measuring growth. The number of students tested in each case ranges from 40 to about 120, with an average of about 75 students. Results are presented for the Total Reading and Total Mathematics subtests.

One unusual feature of the testing procedure should be mentioned. This testing was done by the researchers as part of the project, to obtain the most exact possible information on each child's performance level at the beginning and end of the year. As one feature of that special testing, individual students were assigned to levels of the test which best matched their personal reading level, as reported by the teacher, at the time of testing. Because these students were typically below nominal grade-level in performance, it happened fairly often that students in the upper grades, say grade five,would be tested with a test nominally targeted at grade two or at grade three. This procedure avoided the problem of floor effect on the pretest, which would have worked to invalidate the experiment, since actual gains for many students would occur below the floor and would not be detected.

Thus, although the procedure had advantages for the main purpose of the experiment, it will be seen that it also produced an atypical pattern of standard deviations.

Each line of Table 3 shows the standard deviations for the six grades in one school on one subject. Inspection of the rows indicates that there is a general pattern of fan-spread, for both the reading and the mathematics tests, contrary to the pattern in the publisher's norming data. There is also considerable fluctuation in these standard deviations. However, if all eight standard deviations for the reading test in each grade are averaged, and these averages are compared across grades, the pattern of fan-spread is pronounced. The shift occurs mainly between grades one and two on the one hand, and the four higher grades on the other. A similar general pattern obtains for the mathematics standard deviations. Both of these patterns are shown in the summary lines of Table 3.

The fact that the standard deviations of scores at grades one and two are substantially smaller than in the norm group is as expected, because the data in each row include only students from single schools, deliberately chosen to be relatively low in average performance and restricted in their range of performance. The puzzling aspect of these data therefore is in the size of the standard deviations at the higher grades. These are much larger than would have been expected. The obvious explanation of this pattern is that the procedure of individually assigning test levels did indeed eliminate the "floor" effect on obtained scores, and that this led to the recording of some very low scores. The point to remember about this example is that the low scores are accurate. The distortion occurs not here but in the more usual testing situation, where obtained scores may be restricted in range because of the restrictions of the published test levels.

Further specific analysis of these data could perhaps be pursued, but to do so would lead us away from the general expository aims of this paper. Hopefully, the preceding discussion will serve to indicate that fan-spread is itself an empirically variable phenomenon; it can and should be analyzed in terms of its dependence upon school policies, contexts, and accidental practices of teaching and testing.

The preceding several paragraphs have described how floor effects resulting from test level boundaries can lead to complications involving fan-spread. It should be obvious that similar complications are possible as a result of ceiling effects in other kinds of situations. In the next several paragraphs, we shall consider a different issue -- the consequences of the simultaneous occurrence of fan-spread and the statistical regression effect.

Regression Toward the Mean and Fan-Spread

One serious implication of the fan-spread phenomenon is that it complicates the analysis or comparison of changes or gains over time between two schools or programs. For example, it is not at all unusual to have an experimental program in one school, and a control program in another school, but to have the scores of students in the control group initially higher than those of the experimental school group. This happened, for instance, in the national evaluation study of Project Headstart, and created serious problems for the analysis (cf. Campbell and Erlebacher, 1970). If the two groups were chosen on the basis of some variable which is really associated with their initial ability levels, but not perfectly connected with initial ability, then over time there will be a tendency for the average scores of the two groups to converge. In other words, the initial differences will have had a transitory component, and that component eventually will disappear. This is the well-known

but less well-understood phenomenon of regression to a common mean.

The regression phenomenon is a fully general property of any data set, regardless of what the dependent variable is. In applications of achievement data to program analysis, however, the regression phenomenon can be expected to be operating is a complex way, and to be operating in conjunction with a fan-spread phenomenon which is also in the data. The regression effect will tend to cause the means of the two groups to converge; the fan-spread phenomenon will tend to cause those means to diverge. Thus, these two patterns occurring together can cause compensating movements, and can obscure any real impact of school-to-school differences in a program, or of an experimental treatment.

This difficulty was noted in the Incentives study mentioned earlier which was carried out for the U. S. Office of Education. In that experiment, there were four pairs of schools, each pair including an experimental and control school. For three of the four pairs, the initial score of the control school was higher than that of the experimental school, and these may have been non-accidental differences. This created some ambiguity in interpreting the results of the experiment, because there was some regression effect causing the means of the schools to move toward each other, and some fan-spread causing a tendency for the means to diverge, without taking any account of the experimental treatment.

In that analysis, the only solution was to accept a reduced precision of inference. Fortunately, the school averages initially had been relatively close together, and therefore the numerical magnitude of any fan-spread, or of any regression, had to be small. Furthermore, the two should have tended to cancel each other out. Thus, the conclusion for the purposes of the analysis was that only patterns showing sizeable movement of scores over time would be

interpreted as clear evidence of a treatment impact. For that particular study, the problem was not severe, but it was a hindrance. Dealing with it more directly would have been possible only if prior design work had allowed for these possibilities, and if additional data had been collected (or a strict randomized assignment had been used, so that the inference of equal initial ability would have been tenable).

One Possible Solution -- Compensating Adjustments

As has been indicated, in actual situations fan-spread is found mixed with other complicating sources of influence upon test score behavior, and thus is quite complicated to deal with. This section does not offer practical advice, but outlines one general way that pure fan-spread can be handled. Before describing this approach, it should be repeated from the preceding section that one way to deal conclusively with fan-spread and regression in experimental studies is to make use of strict randomized assignment of subjects to treatments. However, even for experiments, this option usually is not available in field settings, and it is not conceivable for comparisons of existing programs.

A second general approach, not the one to be discussed here, has been suggested by Fennessey (1973) for the comparisons of existing programs. It rests upon a benchmarking and calibration of the scores used by the local agency, so that what is in effect measured by changes in score is a departure from the conditions during the benchmarking period. That approach is basically practical, though inelegant. The suggestions to be made below, though more elegant, have utility primarily as part of a broader effort to analyze test score data quite systematically, and this may not be very practical for school administrators.

To describe the adjustment procedure, let us consider again the artifical

data used earlier. In originally using these data, we assumed that the individual ability levels (70,65,60,55,50 for the High Group, and 50, 45 40,35,30 for the Low Group) were known. The assumption about the dynamics was that progress per unit is proportional to ability level. In actual situations, there is no direct way to know ability levels. Instead, what is typically available is sets of scores collected at, say, two or three time points on a large number of individuals.

Suppose the High and Low Groups of the previous examples instead are School A and School B. Suppose also that an investigator is interested in comparing the "effectiveness" of the two schools. For this purpose, a direct comparison of the average levels of score, or even of changes in levels within a specified time interval, would not be appropriate. For instance, comparing average scores at time point 2, one finds a difference of 120 - 80 = 40 points in favor of School A. Similarly, comparing gains made between time 2 and time 3, one finds that School A is performing better than School B. But because these data from the artificial example were created without any School impact, it is clear that no such impact exists in the data. In other words, the observed score differences arose from differences in initial distribution of ability and from the dynamics of achievement growth.

The dynamics to which we have referred can be summarized by saying that the following equation holds if no other influences are operating:

1. $(\bar{Y}_{i,j+1} - \bar{Y}_{i,j}) = K_j(\bar{Y}_{i,j})$

In this equation, the left side represents the gain observed in an interval, and the equation says that such gain is proportional to initial performance level. To extract a term for the initial time of all (when $j = 1$), we must postulate a value, call it $\bar{Y}_{i,0}$, which is not observed, and regard it as the

ability. Note that we have thus begun to link initial score and ability in the discussion. This is a linkage which in practice is made problematic by many complicating factors. But, for the purposes of this model, admittedly highly simplified, it is satisfactory.

Note that the constant K_j does not depend on which school, i, is being considered. This however, implies that, if no other influences are operating, the following equation should hold for all the schools in a data set at time j:

$$2. \quad \frac{(\bar{Y}_{i,j+1} - \bar{Y}_{i,j})}{\bar{Y}_{i,j}} = K_j$$

Even though K_j may not be known in advance, it can easily be estimated as the average over i of all the left-hand-side expressions in equation 2. Thus, the method provides a means of adjusting the observed gain to remove the influences of the differences in ability between schools. To apply it to the artificial data, the equations are

for School A: $\frac{120 - 60}{60} = 1$

for School B: $\frac{80 - 40}{40} = 1$

$\therefore K_j = 1$ for $j = 1$

Thus, there is no difference between the two schools as far as program impact is concerned.

It can be seen that in this adjustment process, there is no need to deal with the standard deviation. This is because we are now considering not a strange pattern in some outcome measurement data, but instead a formal model of a learning process. The model asserts that individual gain rate is proportional to ability and that ability is measurable by initial score, across individuals, for any time interval, j to j+1. Thus, the fan-spread aspect of the data pattern becomes irrelevant.

General Discussion

Perhaps the more important point here, however, is that the interpretation required the acceptance of some connected assumptions. The conclusions rest on the validity of a particular model as well as on the data. This particular model provides a partial description of how achievement progress is determined. The model used for this example is about as simple as can be imagined. It includes only the influence of ability on progress (neglecting such other possible influences as curriculum, amount of attention, measurement peculiarities, etc., all of which we have had to mention in the practical discussion). Also, it chooses a very simple relation between ability and progress, namely a linear relation. Any models to be used in practical work would have to be considerably more detailed and realistic.

However, even a model as simple as this may serve to raise the usefullness of discussion, by forcing an open dialogue about the exact role of prior ability in influencing gain, and how that role is modified by various educational practices in the concrete situation. Introducing such a model will create debates, and this may be uncomfortable. One important stimulus to the development of models of this general type (e.g., the FEHR-Practicum and related packages) is that without such models interpretation of achievement test data is bound to be crude, and perhaps positively misleading.

References

Campbell, Donald T. Temporal changes in treatment-effect correlations: A quasi-experimental model for institutional records and longitudinal studies, in Proceedings of the 1970 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1971.

Campbell, Donald T. & Albert Erlebacher. How regression artifacts in quasi-experimental evaluation can mistakenly make compensatory education look harmful, in The disadvantaged child (Jerome Hillmuth, ed.), Volume 3 of Compensatory Education, a national debate. New York: Brunner Mazel, Inc., 1970.

Fennessey, James. Using achievement growth to analyze educational programs. Baltimore: Center for Social Organization of Schools, The Johns Hopkins University, Report No. 151, March, 1973.

Planar Corporation. Incentives in education project, impact evaluation report. Washington, D. C.: October, 1972.

Science Research Associates. Evaluating educational growth. Chicago: 1969.